E-041

音声モーフィングのための母音スペクトル間区分線形写像関数自動設計手法

Automatic mapping function designing method modeled by segmental linear function for auditory morphing

高橋　徹† 大西　壮登‡ 森勢　将雅‡ 坂野　秀樹＊ 河原　英紀† 入野　俊夫‡
Toru Takahashi Masato Ohnishi Masanori Morise Hideki Banno Hideki Kawahara Toshio Irino

1. まえがき

本稿では，２つの母音スペクトル間におけるスペクトル距離を最小化する写像関数を自動設計する方法を提案する．音声モーフィングは，２つの音声スペクトルを対数振幅スペクトルの混合によって処理された．その後，周波数軸方向の伸縮が考慮されたモーフィングが行われるようになった．周波数軸方向の伸縮処理は，動的計画法による対応付けることができる．しかし，必ずしも主要な共振周波数同士が対応づけられない点と計算量が多い点が問題である．提案手法は，周波数軸方向の伸縮を考慮した方法であり，母音スペクトル間の主要な共振周波数同士を高速に対応づけることができる．

音声モーフィング[1-6]処理は，音声スペクトルの主要な共振周波数を対応づけた上でパラメタを混合する処理である．対応付けは，スペクトル間の写像関数を設計することに相当し，自動設計が検討されてきた．

ほとんどの場合，写像関数は，スペクトログラムや基本周波数，音声波形，発話コンテキストなどをもとに人手で設計される．この作業には，多くの時間を要し，長時間の音声資料や多数の音声資料のモーフィング音声を生成することを困難にしている．そのため，大量のモーフィング音声を作り評価すること自体が困難である．

人手で写像関数を設計する場合，２つのスペクトル間の主要な共振周波数を対応づけ，隣接する共振周波数間を線形伸縮させる区分線形関数でモデル化してきた．つまり，このモデル化は，区分境界を表すアンカーポイントを配置する問題となる．同様の問題が，音声スタイルマッピング[7,8]や音声テクスチャマッピング[9]処理に含まれており，写像関数の自動設計手法は，他の手法に応用可能である．

本稿では，スペクトル間の対応付けをアンカーポイントの配置問題ととらえ，アンカーポイントを自動配置する方法による写像関数を自動設計する．従来の自動化手法[10]に比べ，主要な共振周波数に自動配置できることを確認した．手動によるアンカーポイントの設定は，数秒の音声にアンカーポイントを設定するために数時間以上かかることがある．しかし，自動化手法の導入により，手動では困難な長い文のモーフィングが可能となる．

スペクトルを動的計画法 (Dynamic Programming)による自動対応付けに比べ，提案法は，約１０分の１の計算量でアンカーポイントを配置することが可能である．計算量を削減できる．

2. 音声モーフィング

本章では，音声モーフィングがどのような分野で利用されているかを述べる．続いて音声モーフィングにおいて，スペクトログラム間の写像関数が果たす役割について述べる．

音声モーフィングは，歌声への連続的表情付与システム[11]，歌唱音声の声質と歌い回し転写技術[12]などの応用システムで用いられている．異なる表情間や声質間の音声資料から，モーフィングによって連続的にスペクトル形状を変え，その量を制御することで表情付けや声質の変化を制御している．音声モーフィングは，人間には発声困難な刺激の連続体を生成できるため，パラメタ表現と知覚の関係を明らかにするツールとして役立つ．能の表現におけるパラメタ表現と知覚の関係を明らかにする試み[13]にも，音声モーフィングが用いられている．これらの研究を効率的に進めるためには，写像関数を自動的に求める方法を確立することが重要である．

２つの音声資料から中間的な音声を生成する音声モーフィングは，音声分析合成の枠組みにおいて，分析によって得られる情報表現を操作・変換し，変換された情報表現から音声を再合成する手法である．音声モーフィングは，拡張された音声分析合成系と位置づけられる．情報表現の具体的な操作・変換とは，２つの音声資料から推定された音源情報と声道情報のパラメタをそれぞれ混合する処理である．しかし，一般に２つの音声資料の発話長は異なり，音源情報や声道情報に対応するパラメタをそのまま混合することはできない．

本稿では，特に音声モーフィングにおける声道情報の写像関数について検討する．時間方向の正規化と周波数方向の正規化を独立に取り扱う．写像関数は，音韻境界や，主要な共振周波数位置を合わせる関数である．発話コンテキストや共振周波数が，写像関数設計の手がかりとなる．

時間方向の正規化の手がかりは，音韻境界を利用できる．音韻境界は，Julian [14]などの大語彙連続音声認識ソフトウェアを用いて音声資料から自動的に求められる．音韻セグメントごとに時間方向に線形伸縮し声道情報を時間方向に正規化することができる．最終的に，周波数方向の正規化を自動化する課題が残る．

一方の時間周波数パラメタを，他方の時間周波数軸上に写像したとき，聴覚的なスペクトル距離が最小となるよう写像関数を設計することが最終的な目的である．提案手法は，聴覚的に最も重要な特徴量の一つである共振周波数を

†和歌山大学システム工学部，‡和歌山大学大学院
＊名城大学理工学部

一致させる写像関数を設計する手法である．本稿では，文献[6]に倣い，主要な共振周波数間を線形伸縮する区分線形関数でモデル化する．区分線形関数は，写像関数を人手で設計する場合，アンカーポイント以外の変位を直感的に把握できることから用いられてきたが，提案手法のモデルは，任意の区分単調連続関数に適応できる． 写像関数設計の自動化は，筆者ら[10]によって進められてきたが，配置したアンカーポイントが共振周波数から外れることがあった．

3. 写像関数の自動設計

３章では，写像関数自動設計の問題点を示し，提案手法について述べる．

3.1 問題点

時間正規化後の２つのスペクトルが与えられると，スペクトル距離の観点で最適な写像関数を設計できる．しかし，DP を用いて全ての時刻において対を成すスペクトル間の写像関数を求めることは，計算量的に現実的ではなく，時間連続性を保つことも困難である．共振周波数位置は，声道形状の変化に応じて変化するが，スペクトルレベルの変化速度より，周波数軸方向の伸縮率変化速度が遅い．写像関数は，時間方向に滑らかに変換する関数である．

写像関数は，共通の周波数空間表現を媒介することで設計する．共通の周波数空間表現は，予め用意された共通の周波数空間表現を代表するスペクトル（テンプレートスペクトル）に写像することで得られる．テンプレートスペクトルは，主要な共振周波数を明示的に表されている．スペクトルをテンプレートスペクトルに DP によって，自動的に写像すると，明示的に表された共振周波数位置に対応付く位置をスペクトルの共振周波数位置とみなせる．与えられた２つのスペクトルを，共通の周波数空間へ写像する関数をそれぞれ求めと，共通の周波数空間を介して写像できる．

この方法により，与えられたスペクトルは，写像先のスペクトルに依存することなく，共通の周波数空間への写像関数を求めることができる． 最終的に，２つのスペクトルが与えられたときの写像関数自動設計問題は，スペクトルを共通の周波数空間への写像関数自動設計問題に帰着する．具体的には，与えられたスペクトルとテンプレートスペクトルとの写像関数設計問題となる．

主要な共振周波数は，概念的にホルマントに対応づけられる．ホルマント推定法を用いて，ホルマントを対応付け，写像関数を自動設計できるようにおもわれる．しかし，この方法には問題がある．例えば，第１ホルマントから第４ホルマントをもとに写像関数を設計する場合を考える．これら４つのホルマントがどのようなスペクトルからでも推定可能ならば，求まったホルマントは一意に対応付き，目的の写像関数もまた一意に設計できる．しかし，ホルマント推定は，ホルマント候補として，いくつかの候補を求めた後に，その中からホルマントらしい候補を選出することによって推定される．そもそも選出候補に現れないこともある．ホルマントが推定できなかった場合には，スペクトル間のホルマントを一意に対応づけられなくなり，写像関数の設計が困難になる． 従って，写像関数を設計するためには，ホルマントとは別の点を対応付ける必要である．

本稿では，予め主要な共振周波数を求めたスペクトルを用い，共通の時間周波数空間を表現する．ただし，主要な共振周波数とは，必ずしもホルマント周波数と一致するものではない．このスペクトルをアンカーポイント付きテンプレートスペクトルと呼ぶ．アンカーポイント付きテンプレートスペクトルと入力スペクトルとを DP 法によりマッチングし，アンカーポイントが対応付いた入力スペクトルの位置を主要な共振周波数とする．それぞれのスペクトルに対して求めた共振周波数を対応付け，隣接する共振周波数間を区分とする区分線形関数でモデル化する方法を提案する．

同一母音の対数振幅スペクトル $S(\omega)$, $X(\omega)$について，次の状況を考える．スペクトル $S(\omega)$の主要な共振周波数は既知であり，スペクトル $X(\omega)$の主要な共振周波数は未知である．この状況において，スペクトル $X(\omega)$の主要な共振周波数を求める．２つの振幅スペクトルを DP 法を用いて，スペクトルを伸縮し，伸縮後に $S(\omega)$の主要な共振周波数が $X(\omega)$上のどの位置に対応付けられたかという情報から，$X(\omega)$のホルマント周波数を知ることができる．しかし，$S(\omega)$, $(0<\omega<\pi)$を N 点で離散化し，DP 処理するために計算量は，N^2 のオーダーとなり効率的ではない．そこで，音声モーフィングが区分線形補間されることから，写像関数のモデルも同様に $S(\omega)$の隣接する主要な共振周波数の間は，線形関数でモデル化することで計算量を IN のオーダーに改善できた [10]．ただし，I はテンプレートに付けたアンカーポイントの数とする．

3.2 自動設計法

前節の方法は，必ずしも $X(\omega)$の共振周波数に対応付かないという問題がある．そこで，$X(\omega)$を全極スペクトルモデルでフィッティングし，得られる共振周波数を $r_1, r_2, ..., r_p$ を予め求める．求める共振周波数の個数 P を十分大きくとると，$r_1, r_2, ..., r_p$ 中に主要な共振周波数がすべて含まれる．$\omega_1, \omega_2, ..., \omega_I$ を $S(\omega)$の主要な共振周波数とする．$\omega'_1, \omega'_2, ..., \omega'_I$ を $r_1, r_2, ..., r_p$ から重複を許さず選ばれ I 個の共振周波数とする．共振周波数の集合を

$$\Omega = \{\omega_0, \omega_1, ..., \omega_I\},$$
$$\Omega' = \{\omega'_0, \omega'_1, ..., \omega'_I\},$$

とする．ただし，区分線形伸縮関数の定義上，便宜的に $\omega_0 = 0$， $\omega_{I+1} = \pi$，とし，$\omega_0 < \omega_1 < ... < \omega_{I+1}$，$\omega'_0 < \omega'_1 < ... < \omega'_{I+1}$，とする．この時，区分線形伸縮関数の区分は ω_i， ω'_i である．区分線形伸縮スペクトル距離を最小とすると主要な共振周波数を対応づけることができる．

$$D(S(\omega), X(\tilde{\omega}) \mid \Omega, \Omega')$$
$$= \int_{-\pi}^{\pi} \{20\log_{10}|S(\omega)| - 20\log_{10}|X(\tilde{\omega})|\}^2 d\omega$$
$$= 2\int_{0}^{\pi} \{20\log_{10}|S(\omega)| - 20\log_{10}|X(\tilde{\omega})|\}^2 d\omega$$
$$= \sum_{i=0}^{I-1} 2\int_{\omega_i}^{\omega_{i+1}} \{20\log_{10}|S(\omega)| - 20\log_{10}|X(\tilde{\omega})|\}^2 d\omega$$

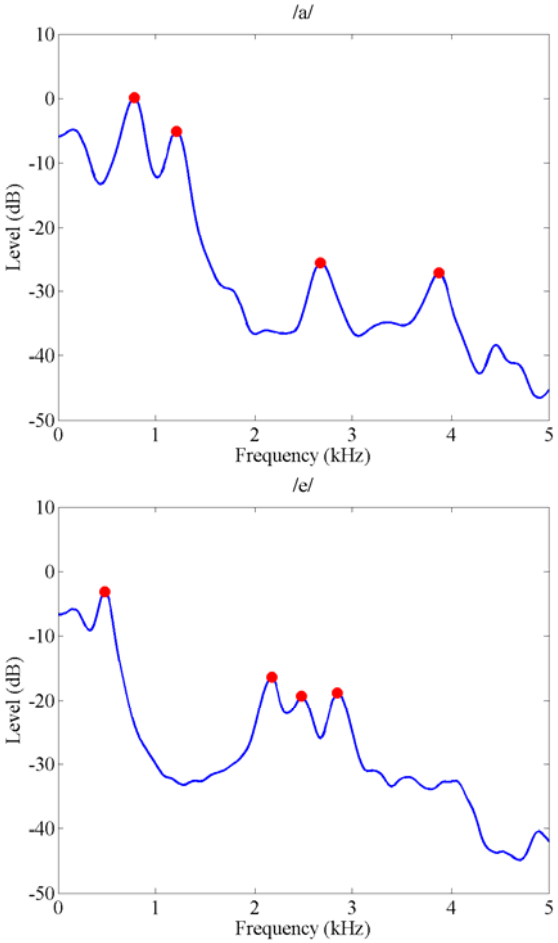

図１　アンカーポイント付きテンプレートスペクトル

$$\tilde{\omega} = \begin{cases} \dfrac{\omega_0' - \omega_1'}{\omega_0 - \omega_1}\omega + \dfrac{\omega_0\omega_1' - \omega_0'\omega_1}{\omega_0 - \omega_1}, & \omega_0 \le \omega \le \omega_1 \\ \dfrac{\omega_i' - \omega_{i+1}'}{\omega_i - \omega_{i+1}}\omega + \dfrac{\omega_i\omega_{i+1}' - \omega_i'\omega_{i+1}}{\omega_i - \omega_{i+1}}, & \omega_i < \omega \le \omega_{i+1} \end{cases}$$

ただし，i=1,2,…,I, とする．

この定義により，スペクトル $S(\omega)$のアンカーポイントは，必ず $X(\omega)$の共振周波数に対応づく．また，計算量は，PI となる．一般にスペクトルの周波数分解能を規定する N に比べ，アンカーポイント数や(I<<N)，主要な共振周波数の候補となる共振周波数の数(P<<N)は十分小さい．

4. 評価実験

声道情報に対応する時間・周波数パラメタ表現の一つにスペクトログラム表現がある．スペクトログラム表現の中でも STRAGIHT[15]スペクトログラムは，ピッチ同期化処理された滑らかなスペクトル表現である．また，音源情報とほぼ完全に分離されたスペクトル表現である．そのためスペクトル操作に対する再合成音声の品質が高い．このことから STRAIGHT は，音声モーフィングに適した音声分析再合成の枠組みであり，提案手法の評価に STRAIGHT スペクトルを用いた．

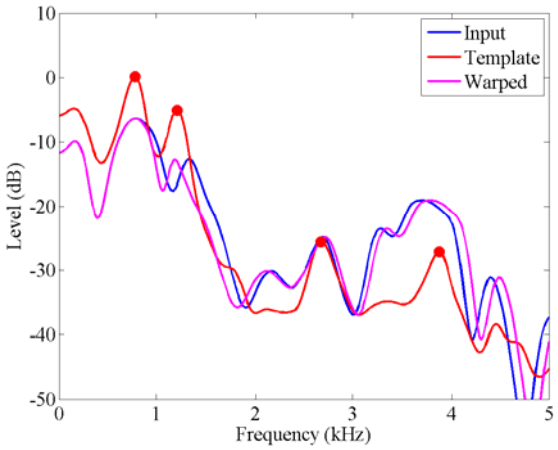

図２　スペクトルの周波数伸縮

自動設計された STRAIGHT スペクトル間の写像関数を評価する．44.1kHz でサンプリングし，16bit 線形量子化された８名の話者のデータベースからアンカーポイント付きテンプレートスペクトルを作成した．図１にアンカーポイント付きのテンプレートスペクトルを示す．図２は，アンカー付きテンプレートスペクトル（赤）のピーク位置に合うように入力スペクトル（青）が周波数軸方向に変形する様子を表している．テンプレートと DP 法で変形量を求め，変形スペクトル（ピンク）を求めると，テンプレートスペクトルのアンカーポイント位置に，入力スペクトルのピーク位置が移っていることを確認できる．

５母音のアンカーポイント付きテンプレートスペクトルを用意し，提案手法と，スペクトル間を DP 法で対応付ける従来方法[9]を比較する． アンカーポイント付きテンプレートスペクトル作成に用いなかった話者で発話内容も異なる音声に対する評価を行う．発話「たのしんでる」を用いて評価を行った．

図３は，STRAIGHT スペクトログラムに，従来法を用いて母音中心時刻のアンカーポイントの対応付く位置を白○で表した図である．白の縦線は，母音中心時刻を表し，左から順に，/a/,/o/,/i/,/e/,/u/に対応している．配置されたアンカーポイントは，主要な共振周波数に対応付けられていない．文献[10]では，主要な共振周波数にアンカーポイントを対応付けられていた．しかし，本実験では， アンカーポイント付きテンプレートスペクトルが，他の話者群から求められたため， アンカーポイントの対応付けが劣化したと考えられる．

図４は，提案法を用いて主要な共振周波数にアンカーポイントを対応付けた結果である．アンカーポイントは，主要な共振周波数に対応付けられている．

実験は，スペクトルを音声帯域に相当する 0～5kHz の範囲で区分線形伸縮を行った．この帯域を 233 点で表した．つまり，N=233, とした．その他 I=4, P=7, とした．この条件では，従来法は，DP のコストを計算するために 932 の格子点に対して距離計算が必要であるのに対して，提案法は，28 の格子点に対する距離計算で済み，高速にアンカーポイントを自動配置できる．

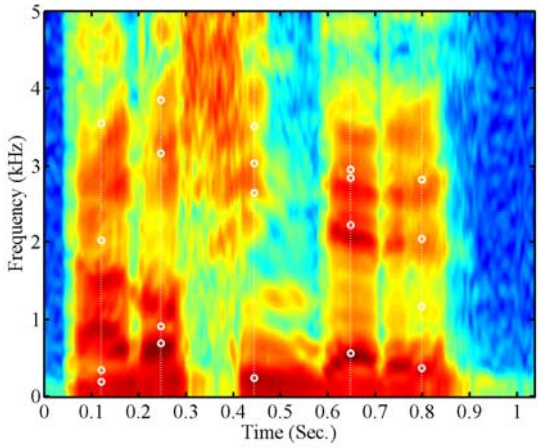

図 3　従来法を用いたアンカーポイントの自動配置

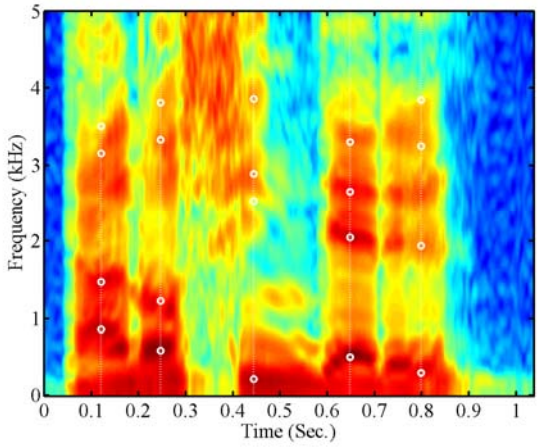

図 4　提案法を用いたアンカーポイントの自動配置

5. まとめ

アンカーポイント付きテンプレートスペクトルを媒介して２つのスペクトル間の写像関数を自動設計する方法を提案した．提案法は，アンカー付きテンプレートスペクトルが，対応付けようとする話者と異なる複数の話者から作成されたにも関わらず，主要な共振周波数にアンカーポイントを配置できた．提案手は，従来法に比べ，計算量を３０分の１に削減できた．従来アンカーポイントは，手動あるいは計算量の多い自動化手法によって配置されていたため，多くのポイントを配置できなかった．計算量が削減できたことから，より多くのポイントを配置でき，スペクトログラム間をより滑らかに写像できる可能性がある．

6. 謝辞

本研究の一部は，文部科学省リーディングプロジェクト e-Society「ユーザ負担のない話者・環境適応性を実現する自然な音声対話処理技術」の支援，科学技術振興機構，戦略的創造研究推進事業 CrestMuse プロジェクト「時系列メディアのデザイン転写技術の開発」の支援により行われた．

[1] X. Serra, ``Sound hybridization techniques based on a deterministic plus stochastic decomposition model,'' *Proc. International Computer Music Conference 1994.*

[2] E. Tellman, L. Haken and B. Holloway, ``Timbre morphing of sound with unequal numbers of features,'' *Journal of AES*, Vol.43, pp.678—689. 1995.

[3] N. Osaka, ``Timbre interpolation of sounds using a sinusoidal model,'' *Proc. International Computer Music Conference* 1995.

[4] M. Slaney, M. Covell and B. Lassiter, ``Automatic audio morphing,'' Proc. International Conference on Acoustics, Speech and Signal Processing 1996, pp.1—4, 1996.

[5] Z. Settle and C. Lippe, ``Real-time audio morphing,'' *Proc. 7th International Symposium on Electronic Art* 1996.

[6] H. Kawahara and H. Matsui, ``Auditory morphing based on an elastic perceptual distance metric in an interference-freq time-frequency representation,'' *Proc. ICASSP 2003*, Vol.I, pp.256—259. 2003.

[7] T. Takahashi, et al., ``Voice and emotional expression transformation based on statisics of vowel parameters in an emotional speech database,'' *Proc. Interspeech 2005*, pp.537—540, 2005.

[8] T. Takahashi, et al., ``Speech style conversion based on the statistics of vowel spectrogram and nonlinear frequency mapping,'' *Proc. EUSIPCO 2006, 2006.*

[9] T. Takahashi, et al., ``General framework for flexible speech style manipulation and synthesis,'' *Proc. WESPAC IX 2006.*

[10] T. Takahashi, ``Automatic assignment of anchoring points on vowel templates for defining correspondence between time-frequency representations of speeches,'' *Proc. Intersppech 2006*, pp.2514—2517, 2006.

[11] 米澤　他，「擬人化ジェスチャ表現を用いた音声への連続的表情付与システム」，日本音響学会誌，Vol.62, No.3, pp.233—243. 2006.

[12] 河原　他，「歌唱モーフィングに基づく声質と歌い回し転写の知覚的検討」，*Proc. Interaction2007*, 2007.

[13] H. Kawahara, ``Voice as artistic expression in Noh,'' *Proc. 4th ASA/ASJ Joint Meeting*, 2006.

[14] hppt://julius.sourceforge.jp/en_index.php?q=en/index.html

[15] H. Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigne', ``Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,'' *Speech Communication*, Vol.27, pp.187—207. 1999.